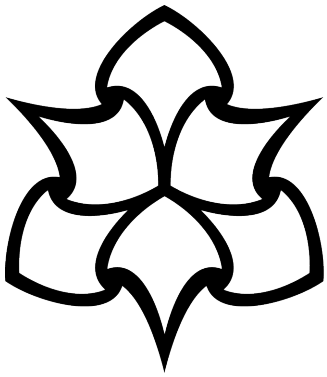
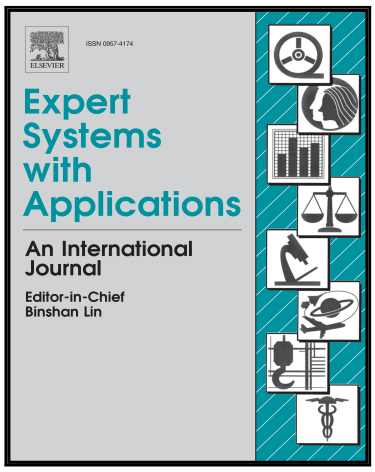




Coarse and fine identification of collusive clique in financial market

Jia Zhai , Yi Cao , Yuan Yao , Xuemei Ding , Yuhua Li

PII: S0957-4174(16)30593-0
DOI: 10.1016/j.eswa.2016.10.051
Reference: ESWA 10956

To appear in: *Expert Systems With Applications*

Received date: 5 June 2016
Revised date: 20 September 2016
Accepted date: 23 October 2016

Please cite this article as: Jia Zhai , Yi Cao , Yuan Yao , Xuemei Ding , Yuhua Li , Coarse and fine identification of collusive clique in financial market, *Expert Systems With Applications* (2016), doi: 10.1016/j.eswa.2016.10.051

Highlights

- We analyse the essential features of collusive clique in financial market.
- We model collusive trading actions by directed and undirected graphs.
- The trader and transaction are transformed and classified into clusters by k-means.
- The clique is identified by a dynamic programming based approach.

Coarse and fine identification of collusive clique in financial market

Jia Zhai,

Salford Business School, University of Salford, 43 Crescent, Salford M5 4WT, UK

j.zhai@salford.ac.uk; jia.zhai1982@gmail.com

Yi Cao,

Division of Mathematics and Computation, School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Minshull House, 47-49 Chorlton St, Manchester M1 3FY, UK

Jason.caoyi@gmail.com ; jason.yicao@outlook.com

Corresponding author: Yuan Yao,

Institute of Management Science and Engineering, Business School, Henan University, 475004, Jinming District, Kaifeng, Henan Province, China

Email: yaoyuan@henu.edu.cn ; prof.yuanyao@gmail.com

Xuemei Ding,

Faculty of Software, Fujian Normal University, Upper 3rd Rd, Cangshan, Fuzhou, Fujian Province, 350108, China

xuemeid@fjnu.edu.cn

Yuhua Li,

School of Computing, Science & Engineering, University of Salford, 43 Crescent, Salford M5 4WT, UK

Y.Li@salford.ac.uk

Abstract

Collusive transactions refer to the activity whereby traders use carefully-designed trade to illegally manipulate the market. They do this by increasing specific trading volumes, thus creating a false impression that a market is more active than it actually is. The traders involved in the collusive transactions are termed as collusive clique. The collusive clique and its activities can cause substantial damage to the market's integrity and attract much attention of the regulators around the world in recent years. Much of the current research focused on the detection based on a number of assumptions of how a normal market behaves. There is, clearly, a lack of effective decision-support tools with which to identify potential collusive clique in a real-life setting. The study

in this paper examined the structures of the traders in all transactions, and proposed two approaches to detect potential collusive clique with their activities. The first approach targeted on the overall collusive trend of the traders. This is particularly useful when regulators seek a general overview of how traders gather together for their transactions. The second approach accurately detected the parcel-passing style collusive transactions on the market through analysing the relations of the traders and transacted volumes. The proposed two approaches, on one hand, provided a complete cover for collusive transaction identifications, which can fulfil the different types of requirements of the regulation, i.e. MiFID II, on the other hand, showed a novel application of well-known computational algorithms on solving real and complex financial problem. The proposed two approaches are evaluated using real financial data drawn from the NYSE and CME group. Experimental results suggested that those approaches successfully identified all primary collusive clique scenarios in all selected datasets and thus showed the effectiveness and stableness of the novel application.

1. Introduction

The identification and prevention of manipulation in the financial markets has been a key focus of academic and regulatory attention in recent years, particularly since the major financial crisis in 2008, and the flash crash of 2010. There are a number of ways the financial markets can be manipulated, which can cause extensive harm to the smooth functioning and integrity of the markets. Trade-based manipulation, whereby manipulation action occurs purely through buying and selling sequences (Franklin & Douglas, Stock Price Manipulation, 1992) is a primary form, and its primary target is the equity price or volume. Manipulations targeting on equity price through single trader's action was extensively covered in our previous work (Cao Y. , Li, Coleman, Belatreche, & McGinnity, 2013) (Cao, Li, Coleman., Belatreche, & McGinnity, 2014) (Cao, Li, Coleman, Belatreche, & McGinnity, 2015) (Zhai, Cao, Yao, Ding, & Li, Sep 2016). Volume-targeted manipulation, which occurs when rogue traders act to give the false impression that a market is subject to a high volume of trading (Franklin & Douglas, Stock Price Manipulation, 1992) (Franklin, Lubomir, & Mei, 2006) was also investigated by (Cao Y. , Li, Coleman, Belatreche, & McGinnity, 2015). However, in practice, another format of market manipulation can be characterised by a group of traders circulating large numbers of shares of equity among themselves in huge numbers of transactions. The result of such actions can be either a created increasing trend of the equity price or a false appearance of active behaviours on the equity trading volume (Palshikar & Bahulkar, 2000). This format is termed as collusive clique based manipulation (CESR, 2011) (Palshikar & Bahulkar, 2000) (Palshikar & Apte, 2008) because this format usually involves a big clique of traders trading together as a collusion to achieve certain effect on the price or volume of a target equity. The buying and selling activity that results from this does not affect the ownership of that stock in the end, but gives the fraudulent impression of intensive and active interest of the stock to market participants (Cumming, Zhan, & Aitken, 2012). In this format, the trading action of each single trader usually seems legitimate, however, the customised trade sequences by the collusive manipulators complies with their manipulative expectations. If any single leg or part of such trades were to be monitored, it would not be identified as collusive trading. An obvious characteristic of the collusive trading is the group of traders, who trade among themselves for misleading the market, and also trade with others for legitimate transactions as well during the period of the manipulation. Such mixed trading actions are another characteristic of collusive trading. Because of their complex and random strategies, it is hard to give an accurate definition of collusive trading. A widely accepted definition is: the collusive clique is the

group who "heavily" trade among themselves than others. The trading behaviours among the clique is collusive trading (Palshikar & Bahulkar, 2000) (Palshikar & Apte, 2008). According to this definition, much of the existing related literature examined the collusive cliques on the basis of similarities in their activity (Wang, Zhou, & Guan, 2012). Their assumption was the traders in collusive cliques acting identically. However, the assumption does not hold in practice when traders usually trade among themselves as well as with others parties. Palshikar et al investigated this topic thoroughly in (Palshikar & Bahulkar, 2000) (Palshikar & Apte, 2008) through a k Nearest Neighbour based algorithm. They assumed each of the traders in a clique trading with all other collusive traders and their crossed transactions can be identified as a feature of collusive clique. This assumption is not true in some cases in practice when the each collusive traders only trade with a small part of a big collusion. Such case was defined in (CESR, 2011) as "passing the parcel", a manipulation format, where a big shares of certain equity are transacted through multiple traders along a certain pathway as a "parcel" without changes of the actual owners. Therefore, the existing literatures were all based on assumptions of how collusive traders behave. There is, clearly, a lack of effective literature that provides a thorough analysis of the inherent features of collusive cliques and a practical identification requirements from regulators. Hence this paper contributed to the literature by seeking to fill this gap and to suggest a novel approach to detecting a wide spectrum of collusive cliques and to providing a decision-support tool for the regulatory departments. Specially, it built upon our previous work (Cao, Li, Coleman, Belatreche, & McGinnity, 2015) and provided a thorough discussion and analysis of all appropriate scenarios, and identified and quantified the key features thus aroused. In light of this, a clear formulation of the problem is given, along with explanation of the relevance of conceptual models. As far as we are aware, this paper presented the first study that solved the collusive clique detection problem in a practical way through two proposed approaches based on popular data mining algorithms, which also shows an effective application: a challenging problem in real financial area, can be extracted and formulated to a standard format and then solved by applying tailor-made popular data mining algorithms.

The first proposed approach was to generally identify the collusive trend of all traders in collected trading records. We termed it as: coarse identification of the collusive clique. This approach was to group all traders into a number of clusters, where the traders tend to trade heavier with traders within the cluster than the ones outside. We proposed a transformation of the transaction data together with *k*-means clustering algorithm to group the traders according to their trading behaviours. The result of this approach aimed to show a general tendency of how traders gather together for transactions. This was especially useful when experiencing high-frequency trading (HFT), where the characteristics of whole transactions were important instead of every single one among the huge number of orders and trades generated by HFT. In this approach, the 'distance' between any two traders was defined as the transacted volume: the larger the volume, the closer the 'distance'. Through this, the tailor-made *k*-means clustering approach can be easily applied with reduced time complexity since the calculation of 'distances' among 'points' in standard *k*-means has been moved to the proposed transformation of the transaction data. This showed an effective application of well-known data mining algorithm on an unsolved and challenging problem.

The second proposed approach, on the other hand, was a fine identification, which was to accurately identify the "passing the parcel" transactions that were intendedly created by collusive traders. This was particularly useful when the regulatory monitoring intends to find out in detail which traders are involved in collusive transactions and how they organize such activities. This fine identification was formulated as a simplified Knapsack problem, a well-known combinatorial optimisation problem. In this approach, transaction records were considered as the knapsack and the symbolic sum of traders in the transactions were considered as the weight of the knapsack while the number of the transactions shall be maximized. To solve this, a tailor-made unified dynamic programming based algorithm (DP) was proposed. DP is usually used in complex optimization problems. In our approach, DP was proposed as a recursive problem, where each possible solution on current transaction can be optimized by selecting either the solution based on last transaction or the one only based on the current transaction. Through this, the fine identification problem was split into several smaller problems. This idea is popular in complex optimization problems where the original problem is usually split into a collection of simpler sub-problems. In addition, to make the problem more general, margins were included in the optimization process to compensate the uncertainty occurred in real trading scenarios. Therefore, this approach also showed that the traditional optimization algorithm, which was usually applied in scheduling and traveling, can also be revised and applied to effectively solve complex financial problems.

Therefore, two computational approaches, one for generally collective trend, one for accurate activities, gave a complete cover for collusive transaction identifications, which can fulfil the different types of requirements of the regulators. The recognition of the collusion with heavily transactions or suspiciously "parcel passing" transactions led to a detection of potential collusive clique. The validity of those two proposed approaches have been assessed through extensive experiments, carried out using real data from the different financial products, which showed effectiveness in collusive clique and has yielded a considerable body of evidence in support of the proposed decision-support expert systems for the regulators.

The remainder of this paper takes the following form. Section 2 gives a brief overview of wash trade manipulation and detection. Section 3 contains analysis, formulation and description of wash trade features, and of the proposed approach to detection. Section 4 gives an evaluation of the performance of the proposed approach and Section 5 gives a conclusion.

2. Collusive clique and its Detection

2.1 Trading in Collusive Clique

Traders in capital markets use limit or market orders to buy or sell volumes of an equity, at a specific or better price (a better price, in this context, being a higher selling or lower buying price) (SEC, 2011). When orders are linked under order-matching rules, the transaction takes place. In most exchange markets, the transactions have been recorded across the time as the examples in Table 1, where five traders, labelled A to E, sell and buy the same equity among each other. Following the illustrations in (Palshikar & Apte, 2008), we represent the transactions records in Table 1 as a directed graph in Figure 1, where each trader is represented as a node linked by the arrows representing the transaction directions and the numbers on the arrow represent the transacted

volumes. From Figure 1, we can observe that the total volume between E and A, B, C and D is merely 150 shares, while even the minimal total volume of transactions among the group A, B, C, and D is the 290 shares between D and A, B and C. It is obviously that the highlighted traders A, B, C, and D trade heavier among themselves than with E, who has only two transactions with A and D with total volumes 150. From the intuitive observation, we can generally identify collusive trend among A, B, C, and D based on the transaction records in Table 1. At the same time, we can also observe a transaction loop among traders A, B, and C from the highlighted arrows in Figure 1. The loop is composed of five transactions, highlighted in Table 1. The transaction #1 and 3 are all from trader A to B with volume 300 and 250 respectively. We combine those two transactions with same seller and buyer together as the aggregated transaction with the volume 300+250=550. Similarly, the transaction #2 and 4 can also be aggregated as transaction with volume 450+100=550. By this, we find a directional loop: A selling 550 shares to B, B selling 550 shares to C and C selling back 550 shares to A. Such transaction loop has been done within 20 minutes. Although we cannot establish an illegitimate conclusion based on this, identifying the occurrence of such transaction loop within a short time period is still an important task in monitoring trading behaviours on financial market.

Although each single transaction in a collusive clique following the matching rules, unlike legitimate ones, the collusive transactions exhibit what the Financial Conduct Authority (FCA) describes as "no change in beneficial interest or market risk," or "the transfer of beneficial interest or market risk only between parties acting in concert or collusion, other than for legitimate reasons" (FSA, 2006). The Committee of European Securities Regulators (CESR) extends this description, calling collusive clique a “deliberate arrangement in concert or collusion” (CESR, 2011). The Chicago Mercantile Exchange (CME) formalised a new rule on 28 August 2014. This rule, known as Rule 575, was subsequently adopted by the US Commodity Futures Trading Commission (CFTC) (CME, 2014) and states that “no person shall enter messages to the market as pre-arranged collusion with intent to mislead other participants”. While such actions mean that the regulatory bodies (the FCA, CESR and CFTC) have clearly defined collusive clique and their trading activities, they have yet to provide or even suggest any quantitative method for detecting it.

Table 1. Transaction sequences

Transaction #	Seller	Buyer	Time	Price	Volume
01	A	B	09:00:000	125.5	300
02	B	C	09:05:000	124.9	450
03	A	B	09:05:100	125.5	250
04	B	C	09:07:000	125.2	100
05	B	A	09:10:000	125.3	50
06	C	A	09:11:000	125.3	550
07	A	E	09:13:000	125.3	100
08	E	D	09:13:100	125.3	50
09	D	B	09:14:000	125.3	150
10	A	D	09:16:100	125.4	40
11	D	C	09:17:100	125.4	100

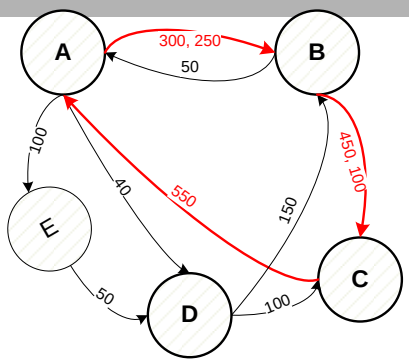


Figure 1 The directed graph illustration of transaction records in Table 1

Table 1 gives an example of the simplest form of the collusive clique, which can be easily observed through the illustration in Figure 1. Collusive clique in practice involves a big number of transactions and traders, which make it hard to identify through simple graph representations and observations. However, the basic structure and strategy in complex collusive cliques are identical to the simple examples in Table 1, which can be summarized as:

- For general collusive trends (coarse identification), the clique is the group, where the volumes of transactions inside the group are significantly larger than outside. The “larger” is a relative indicator and variate across time and different traders.
- For accurate parcel passing transactions (fine identification), the loop starts and ends at the same trader and the transaction volume between any two traders are identical (To avoid detection, clever manipulators ensure that transaction volumes are close but not completely identical.).

2.2 The detection of Collusive Clique

As far as the authors are aware, there are no extant studies that focus on detection of collusive clique and their activities in capital markets. There is some related work based on shared trading behaviours (i.e. the buying and selling of equities in similar ways). A spectral clustering approach has been developed (Franke, Hoser, & Schröder, 2008), whereby a trading behavioural-based network is created and any behaviour that deviates from that network's norm is reported as irregular; underpinning this is the assumption that a trader's current behaviours echo those of his/her previous trading network. A graph clustering algorithm, intended to detect groups of colluding traders, has been suggested (Palshikar & Apte, 2008), using a stock flow graph to elucidate the relationship between traders, whereby those with "heavy trading" in their network are clustered as collusive. A recent work (Wang, Zhou, & Guan, 2012) presents a new approach to detecting collusion in trading, and in this the correlation matrix for a single trading day is given. In this, trader behaviour is shown in the form of an aggregated time series of signed volumes of submitted orders. Pearson's product-moment coefficient is used to quantify the similarities in trader behaviour shared by multiple individual traders, and those cliques showing a coefficient above a user-specified threshold are designated as “suspicious collusions”. The study presents experimental data, drawn from the evaluation of order data for futures traded on the Shanghai Futures Exchange. Order volumes and directions (buy/sell) constitute the “signed order volume”, and price information is ignored on the grounds that order prices do not affect what traders do (Wang, Zhou, & Guan, 2012). The market, however, is considerably affected by order price, as the market impact measure makes clear (Hautsch &

Huang, The market impact of a limit order, 2012), thus the market moves generated by traders' actions (in the form of orders) are fundamental to transactional costs (ITG, 2010). Thus to ignore price information is inappropriate - price information distinguishes the intentions of traders, and is a core element of wash trade manipulations. In mid-2011, a technique was released, that had been developed by the CME to arrest wash trade activity at the "engine level" (Patterson, Strasburg, & Trindle, 2013). This was further updated in the summer of 2013 (Bowen, 2013). Unfortunately, this approach restricted itself to monitoring identically-priced buy/sell orders emanating from trading accounts that shared beneficial ownerships. The consequent inability of this approach to monitor collusive trading involving multiple orders or traders left the way clear for colluding parties to fabricate transactions and thus generate a false impression of (generous) trading volumes. Nonetheless, regulators have been visibly pursuing those involved in collusive clique activities. In December 2012, the Securities and Exchange Commission of Pakistan investigated a collusive trading case (Jamal, 2012), while in March 2013, US regulators inspected the work of traders who had been acting as both buyer and seller in the same transaction. In the wake of this investigation, the authorities claimed that, potentially, several hundred collusive trading actions occur daily on the CME and Intercontinental Exchange (ICE). In June 2012, the regulatory financial authorities in Hong Kong declared that attempts to enter into “collusive” or “matched” trading constitute crimes of financial manipulation, regardless of whether that activity has had, was intended to have, or could have had, the effect of misleading market observers (Loh & Cumming, 2012). Despite a lack of effective detection tools, this declaration marked an important line in the sand; any attempts at collusive or matched trading are now openly acknowledged as financial crimes. So far, academic research has tended to define analogous behaviours, with a view to detecting overall trading collusions. This does not permit a precise or definite result, but can show correlations between trading activities and defined clusters of traders. Meanwhile, industry-generated detection/monitoring approaches have been restricted to simple collusive formats and are thus easily bypassed by simple changes on the part of manipulative traders. An interesting point is that little seems to have been done in the way of analysing strategic behaviour in collusive clique and its trading behaviours as a whole as well as in individual, or of developing a detection method that takes into account the complete range of tactics that may be used by collusive traders, but the regulatory authorities urge to design and implement appropriate tool for improving and speeding up the decision making in the collusive clique detection process. This is a clear gap in the research and understanding of this field, which the current paper seeks to address.

3. Collusive Clique Detection Methodology

3.1 Terminologies used in analysis

In this paper, the analysis of strategic behaviours involved in collusive clique follows the terminologies given in (Tsang, Olsen, & Masry, 2013), with some revision. The conclusion of a collusive transaction loop is illustrated by the aggregated position of the entire colluding group; in this context, “position” is the volume of equities that a trader holds. Since collusive transaction loop comprises fraudulent, rather than genuine, trading actions, the ultimate position of each participating (colluding) trader tends to remain unchanged, in order to minimise financial loss. Thus the position of the group as a whole remains, in the end, virtually unchanged. In the course of the collusive transaction loop, positional change occurs as a result of orders from collusive traders, and can be defined as follows:

Position + Transaction → Position

A sequence of transaction combine to form position, thus:

$$\text{Position} = \{(\text{Tran}_1), (\text{Tran}_2)\ldots (\text{Tran}_n)\},$$

and each transaction is defined in the following scalar form:

$$\text{Tran} = \{\{\text{traders involved in this transaction}\}, \text{Time}, \text{Price}, \text{Volume}\}.$$

In the scalar form, we only record the ID of the traders involved in this transaction (usually two traders) with no information reflecting the selling and buying. We also represent the transaction with its direction in a vector form. Following the terminology in (Tsang, Olsen, & Masry, 2013), the selling and buying actions are represented by negative and positive signs affixing to the trader ID and Volume respectively. The transaction vector is defined as:

$$\overrightarrow{Tran} = \{-\text{Seller\_ID}, \text{Buyer\_ID}, \text{Time}, \text{Price}, \pm\text{Volume}\}.$$

Using the vector format, we can re-write the transactions #1, 3 and 5 in Table 1 as:

$$\overrightarrow{Tran_1} = \{-\text{A}, \text{B}, 09{:}00{:}000, 125.5, 300\}$$

$$\overrightarrow{Tran_3} = \{-\text{A}, \text{B}, 09{:}05{:}100, 125.5, 250\}$$

$$\overrightarrow{Tran_5} = \{\text{A}, -\text{B}, 09{:}10{:}000, 125.3, 50\}$$

If switching the sell and buyer position in $\overrightarrow{Tran_5}$, we can simply re-write the vector as

$$\overrightarrow{Tran_5} = \{-\text{A}, \text{B}, 09{:}10{:}000, 125.3, -50\},$$

which shows that in theory, B selling to A 50 shares equals to A selling to B -50 shares.

Based on the scalar and vector forms, we further define the transaction aggregations. For the scalar form, if the traders involved in two transactions are the same, we aggregate the two transactions by only accumulating the transacted shares. Therefore, the transactions #1, 3 and 5 in Table 1 can be represented and aggregated in scalar form as:

$$\text{Tran}_1 = \{\{\text{A}, \text{B}\}, 09{:}00{:}000, 125.5, 300\}$$

$$\text{Tran}_3 = \{\{\text{A}, \text{B}\}, 09{:}05{:}100, 125.5, 250\}$$

$$\text{Tran}_5 = \{\{\text{A}, \text{B}\}, 09{:}10{:}000, 125.3, 50\}$$

$$\text{Tran}_1 + \text{Tran}_3 + \text{Tran}_5 = \{\{\text{A}, \text{B}\}, -, -, 600\}$$

The time and price information are ignored in scalar form transaction aggregation since the coarse identification targets on merely the traders' collusive trends, which can be shown according to their transaction "heaviness" among themselves. The "heaviness" is usually defined as the total aggregated transacted volumes (Palshikar & Apte, 2008) (Wang, Zhou, & Guan, 2012). Based on this, we can aggregate corresponding transactions together in Table 1 and illustrate in an undirected graph (since the transactions are in scalar format) as Figure 2. From this figure, we can observe the collusive trend clearer as the highlighted nodes.

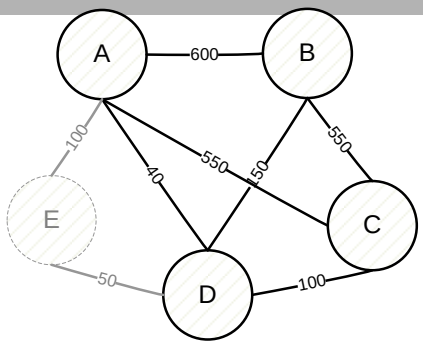


Figure 2 The undirected graph illustration of aggregated transaction volumes in Table 1

Similarly, we can also define the aggregation of the transactions in vector forms the same way as the scalar form: if the traders in two vector forms are the same (including the sign), we aggregate the two transactions by summing up the signed values of the transacted shares; in addition, if the traders in two transactions are different ('–A' is different from 'A'), we aggregate the trader ID by symbolic calculation, i.e. '–A'+'A'='0', and record the shares in each transaction as a list. By this, we can aggregate the vector forms of transactions #1, 3 and 5 in Table 1 as

$$\overrightarrow{Tran_1} + \overrightarrow{Tran_3} + \overrightarrow{Tran_5} = \{-\mathrm{A} - \mathrm{A} - \mathrm{A}, \mathrm{B} + \mathrm{B} + \mathrm{B}, -, -, 300 + 250 - 50\} = \{-3 * \mathrm{A}, 3 * \mathrm{B}, -, -, 500\}.$$

Apparently, the vector form aggregation shows the net transacted volumes between trader A and B. The actual result of transactions #1, 3 and 5 is A selling B 500 shares through 3 times transactions. Thus, if we calculate the aggregated transaction of #1-6, we can have:

$$\sum_{i=1}^{6} \overrightarrow{Tran_i} = \{-\mathrm{A}, \mathrm{B}, -, -, 500\} + \{-\mathrm{B}, \mathrm{C}, -, -, 550\} + \{-C, A, -, -, 550\} = \{'0', -, -, \{500,550,550\}\}$$

The trader ID is summed up as '0' by symbolic calculation and the three transaction volumes are recorded as a list $\{500,550,550\}$. The zero-valued symbolic sum of trader ID, which suggests that each trader of a group is acting at both the buying and selling ends of the market, together with the similar value of three transacted shares show that after a sequence of similar sized transactions, almost no genuine equity shares change hands. Thus the calculated results strongly suggest that traders A, B, and C trade suspiciously as "passing the parcel". We can also illustrate the aggregated the vector forms of transactions in Table 1 in directed graph as Figure 3. The parcel-passing transactions are highlighted.

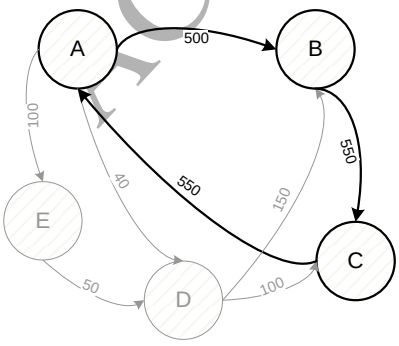


Figure 3 The directed graph illustration of net transacted volumes in Table 1

3.2 Coarse Identification of the Collusive Clique

As discussed in Section 2, financial market can be overloaded by the high-frequency trading (HFT) activities. The HFT traders strategically use millions numbers of orders and transactions to overload, delay and disrupt the markets and other participants. In such cases, analysing every single trade is neither efficient nor contributing to a systematic understanding of the collusive groups of HFT traders. Instead, aggregating the corresponding transactions and identifying their trend of colluding parties and trading actions are important to the regulator to prohibit such disruptive activities (CME, 2014). We define the term for such identification as "coarse identification of the collusive clique" (short as: Coarse Identification).

In coarse identification, we following the definition of "heaviness" of the trading between two traders in (Palshikar & Apte, 2008) (Wang, Zhou, & Guan, 2012) as the total aggregated absolute transacted volumes between those traders. This has been defined as the aggregation of transactions in the scalar form in Section 3.1. As an example, we aggregate the transactions in Table 1 and show the results in Table 2. From the results and the corresponding Figure 2 in Section 3.1, we can observe that the heavier the transactions between two traders, the more likely they are collusive. To further formulate the question, we define a simple transformation function as,

$$f\left(V_{i,j}^{a}\right) = e^{-W\frac{V_{i,j}^{a}}{\sum_j V_{i,j}^{a}}}$$

where $V_{i,j}^{a}$ is defined as the aggregated transactions between trader *i* and *j*, $\sum_j V_{i,j}^{a}$ is the total transacted volumes between *i* and all other traders, and $\frac{V_{i,j}^{a}}{\sum_j V_{i,j}^{a}}$ shows the ratio of volumes between *i* and *j* to total transacted volumes of trader *i*. In this function, we choose $W = 5$ as a scaling weight to transform the volume ratio (ranged from 0 to 1) to its exponential value (ranged from 1 to 0) so that the higher the ratio is the smaller the transformed value is. By applying this function to the transactions in Table 2, we have the results in Table 3 and Figure 4. Considering each trader as a node as the illustration in Figure 2 and the values in Table 3 as the "distance" among the nodes, the "heaviness" among the trader has then transformed to the "distance" between the traders. That is, if traders are very close in distance, they are very likely to be in a cluster. This is basically identical to the traditional clustering problem in data mining area. Therefore, we can summarize the coarse identification of the collusive clique as: given *N* traders with their large numbers of trading records, we aim to partition the *N* traders into *k* clusters, so as to minimize the within-cluster sum of the "distance" of each trader in the cluster to the *k* centre traders. By this, the coarse identification has been formulated as a traditional *k*-means clustering problem.

Table 2 Aggregated volumes of scalar form transactions in Table 1

	A	B	C	D	E
A	-	600	550	40	100
B	600	-	550	150	0
C	550	550	-	100	0
D	40	150	100	-	50
E	100	0	0	50	-

Table 3 Transformed value of the aggregated volumes in Table 2

	A	B	C	D	E
A	-	0.0977	0.1186	0.8564	0.6787
B	0.0995	-	0.1206	0.5616	1.0000
C	0.1011	0.1011	-	0.6592	1.0000
D	0.5553	0.1102	0.2298	-	0.4794
E	0.0357	1.0000	1.0000	0.1889	-

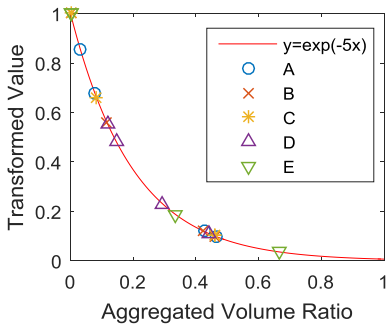
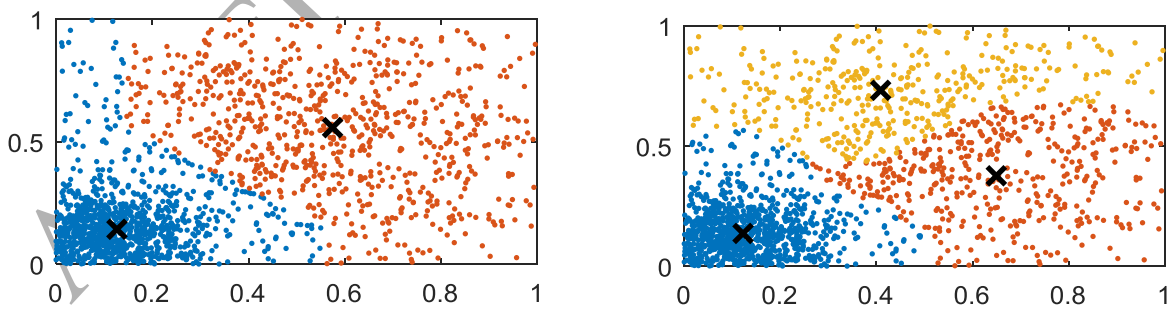


Figure 4 Transformed value of the aggregated volumes in Table 2. A-E represent the traders.

On May 31, 2013, large numbers of collusive trading activities from HFT traders were reported by NANEX (NANEX, Chicago PMI, 2013). Around 550,000 shares of SPY shares trade were generated from a big number of trading accounts within less than 1 second. The regulator noticed this and investigated the HFT traders' collusive behaviours. We applied our proposed coarse identification approach on this example and identified the same collusive clusters of HFT traders as the regulators and NANEX reported. The clusters of traders were illustrated as Figure 5. The dots on the figure represent the traders involved in those transactions. Since we do not use any x-y axis to represent a trader, the representation on Figure 5 is merely to show the relative distances among the traders rather than the exact locations of them. We tried k=2 and 3 in coarse identification approach and both configurations identified the collusive cluster of traders, which are shown as the blue dots in Figure 5(a) and (b).



(a) (b)

Figure 5 clusters identified by coarse identification approach in the HFT examples on May 31, 2013 reported by NANEX, (a) the identified clusters if targeting on 2 collusive groups; (b) the identified clusters if targeting on 3 collusive groups. The left-bottom blue dots represent the collusive HFT traders in the examples.

3.3 Fine Identification of the Collusive Transactions

3.3.1 Formulation of the problem

The second proposed approach we proposed, on the other hand, is a fine identification, which is to accurately identify the collusive transactions such as the format of "passing the parcel" that are intended created by a group of traders. The FCA and CESR have both stated, in consultation reports (FSA, 2006) (CESR, 2011), that accurately identifying collusive transactions is a challenge, since the form that such trading takes may vary, and collusive actions can easily be hidden amidst the huge volumes of normal trading activities. This is further illustrated in report by Nanex, dated 31 May 2013 (NANEX, Chicago PMI, 2013), which shows complex networks and illustrates these using vertices to represent traders and directional connections between them showing the transactions they share. In this paper, we use a directed graph (as Figure 3) with net transacted volumes to represent the potentially collusive transactions. In the directed graph, nodes are used to represent traders, while the short arrows attached to each node represent the net transacted volumes from one trader to another. The arrow direction shows the net transaction flow. As the example shown in Figure 3, the highlighted arrows connect a small group of traders, A, B and C, and the direction of the arrowhead shows the direction of the transacted shares, for example A pointing to B means that trader A sells shares to trader B and thus transfers the equity shares. Therefore, in Figure 3, traders A, B and C collude in a sequence of parcel-passing transactions as a closed cycle. Transactions among them flow in a single direction, either clockwise or anti-clockwise in a trading form of "passing the parcel" (Aitken, Harris, & Ji, 2009). In this context, when a transaction loop is completed, there has been no transfer of beneficial interest across the group, and no one of the collusive traders is in a position any different to that they were in when the process began. In the example of the aggregation of the transaction of #1-6 discussed in Section 3.1, the zero-valued symbolic sum of trader ID ('0') and the similar-valued transacted shares ({500,550,550}) indicates the involved traders passing similar-sized "parcels" through themselves with almost no genuine equity shares change hands. However, in practice, "smart" manipulators intend to mix the collusive trades with other trades to make the sum of trader ID non zero and the transacted share list not similar-valued. As a result, such mixed transactions are easily avoid alerting traditional regulatory inspections (Aitken, Harris, & Ji, 2009).

To compensate the "smart" manipulator's strategy and effectively identify the potential collusive transaction, we propose a straight forward approach with margins of trader (Δid) and transacted shares (Δv). The traders in collusive transactions may not 100% pass the parcel one by one to construct as a loop. The "passing" can have a complex format, where each trader may or may not complete a buy together with a sell transaction. The transacted shares may not be similar as well. As a result, the aggregation of such collusive transactions may have non-zero sum value of trader ID and non-similar transacted shares. Based on our analysis, we can reveal two features of the general strategy used in collusive transactions, which are:

Feature 1: Mostly matched traders. To manipulate the market trading volume, the traders transact among themselves for creating the false appearance only. No one in the collusion like to sustain any loss from this. As a result, most of the traders carry out both selling and buying identical shares through different transactions to maintain their position unchanged. However, to bypass the regulatory monitoring, the collusion might intend to

mix some legitimate transactions with the collusive ones. Thus, not every trader maintain zero position changes in those transactions. If representing those transaction in the vector format and aggregating those traders, we might not achieve a '0' result, which means, additional to a selling and buying pairs, some trader ID may also be involved in intended legitimate transactions. Therefore, we consider a margins of trader ($\Delta\mathrm{id}$) when identifying the traders collusions. That is, if the number of trader IDs in the aggregated result of all signed trader ID is within the margin $\Delta\mathrm{id}$, we consider this as suspicious and carry out the further investigations.

Feature 2: Mostly identical net transacted shares. Similar to the feature 1, traders in the collusion tend to maintain their positions unchanged by both selling and buying identical shares through different transactions. However, they might also transact slightly different to avoid to be recognized as manipulation. To compensate this, we also consider a margin of shares ($\Delta\mathrm{v}$) when monitoring the transacted volumes. When the shares in the transactions are all within the margin $\Delta\mathrm{v}$, we consider it as a potentially collusive transaction.

Thus, we propose a three-step approach for identifying the parcel-passing type collusive transactions, namely:

Step 1: aggregate the transactions with same seller and buyers in the vector format defined in Section 3.1. By this, we can obtain the net transactions shares with directions between any two traders. After this step, all transactions are represented as: $\overrightarrow{Tran_\iota} = \{-\mathrm{Seller\_ID}, \mathrm{Buyer\_ID}, \mathrm{Time}, \mathrm{Price}, \pm\mathrm{Volume}\}$

Step 2: identify a group of transactions $\overrightarrow{Tran_\iota}$, where the number of remained trader IDs after further aggregating all signed trader ID in this group is within the margin $\Delta\mathrm{id}$. After this step, the transaction volumes are recorded as a list in the aggregated result of this group.

Step 3: identify the deviation degree of the transaction volume list. If the volumes are within the margin of shares ($\Delta\mathrm{v}$) (not significantly deviated), we consider this as an identification result: a group of traders transacted collusively in a parcel-passing style.

3.3.2 Fine identification approach

The fine identification approach proposed in this paper is applied to transaction streams in the market. This is to uncover any potential trend of collusive trade in early stage, thus fulfilling the requirements of the latest regulations. An occurrence of a transaction updates the transaction stream. A transaction, as shown in Table 1, comprises a transaction ID, seller and buyer ID, time, price and volume. For the purposes of this study, we assume that transactions in the stream focus on specific, single, financial product and so any product information within the stream can be ignored once that specific financial product is determined. While on one hand this assumption narrows the scope of the study, it has the virtue of conforming to the practicalities of a trading environment, and thus the algorithm can easily be applied to selected product in a genuine trading platform context. As discussed in Section 3.3.1, the approach is composed of three steps. In order for these steps to be taken, the transaction stream must be pre-organised. A physical time sliding window, of size θ_T, is specified and the transactions with same seller and buyers are aggregated in the vector format defined in Section 3.1. After the aggregation, the transaction are in a queue, Q^a, which also maintains a length of θ_T. Thus, if a new transaction $\overrightarrow{Tran_i}$ is coming and the length of the updated queue exceeds θ_T, the earliest transaction should

be popped in order to maintain the length of the sliding window so that the corresponding transactions are re-aggregated again. Since the transaction stream is measured in terms of "event time", calculation of the difference between the time stamps of the first and last transactions in the queue is used to maintain θ_T. Thus, the number of transactions in each queue depends upon the underlying frequency of transactions, and so fluctuates over time.

After the step 1 aggregation, the step 2 is to find out a group of transactions $\overrightarrow{Tran_i}$, where the sum of the signed trader IDs in the group is within the margin Δid. Considering the sliding window θ_T, the step 2 operates on the basis that, among all aggregated transactions $\overrightarrow{Tran_i}$ in θ_T, find transaction sub-group where the trader ID can be calculated and cancelled so that the remained ID number is within the margin Δid. We define the step 2 as: **ID_MATCH**(Q^a), where $Q^a$is the set of aggregated transactions in vector format. Based on our discussion, the function **ID_MATCH**(Q^a) is thus defined as: given a batch of aggregated transactions in Q^a, identify subsets S of transactions from Q^aso that $\left|\frac{\langle\sum_{i\in S} Tran_i\rangle_T}{\sum_{i\in S}\langle Tran_i\rangle_T}\right| \leq \Delta\text{id}$, where operator $\langle\cdot\rangle_T$ represents the number of the traders in the transaction so that $\langle\sum_{i\in S} Tran_i\rangle_T$ represents the trader number after summing up all signed traders and $\sum_{i\in S}\langle Tran_i\rangle_T$ is the number of traders in the group Q^a. Therefore, if group Q^a contains 100 traders and we choose $\Delta\text{id} = 10\%$, it means that once the trader number in sub-group S equal or smaller than 10, we need to further investigate the transactions in S.

If the number of transactions contained within subset S is given as n_s (n_s is equal or smaller in size than Q^a), fundamentally, function **ID_MATCH**(Q^a) can be considered as a simplified version of the Knapsack Problem (Andonov, Poirriez, & Rajopadhye, 2000) (Poirriez, Yanev, & Andonov, 2009) (Zukerman, Jia, Neame, & Woeginger, 2001). The Knapsack Problem refers to the challenge of filling a knapsack, which has capacity W, using a subset of m items $\{1, \ldots, m\}$, each of which has both mass and value, in such a way that the total weight of the combined selected items is the same or less than W, and the highest possible total value is achieved. The trader ID matching problem is a simpler version and can be stated as: given a margin Δid (which is equivalent to the knapsack's size) and a set of items Q^a, each being of (non-negative) a pair of signed trader ID, identify all possible subsets S of items, ultimately to make:

$$\left|\frac{\langle\sum_{i\in S} Tran_i\rangle_T}{\sum_{i\in Q^a}\langle Tran_i\rangle_T}\right| \leq \Delta\text{id}.$$

Given the similarity of those two problems, the approach often used to solve the Knapsack Problem, i.e. dynamic programming, is used in **ID_MATCH**(Q^a). Dynamic programming requires a number of sub-problems, so that the solution to each can be found using "smaller" sub-problems, and thus the original problem can be solved easily once all sub-problems have been resolved (Kleinberg & Tardos, 2005). Dynamic programming has been thoroughly studied in optimisation problems in (Ni, He, Wen, & Xu, 2013) (Jiang & Jiang, 2014). Thus, the intention is to solve this simplified form of the Knapsack Problem under N transactions, and Δid, and to denote $\sum_{i\in Q^a}\langle Tran_i\rangle_T$ as T_{Q^a} and the final subset of transactions in the optimum solution to the original problem as S_N. We define **OPT_ID**$\left(N, T_{Q^a}, \Delta\text{id}\right)$ to represent the sum of the trader IDs of the first N

transactions in subset S, under constraint $\left|\frac{\mathbf{OPT\_ID}\left(N,T_{Q^a},\Delta\mathrm{id}\right)}{T_S}\right| \leq \Delta\mathrm{id}$. Therefore the trader sum of the first $N-1, N-2, \ldots, 1$ orders can be shown as $\mathbf{OPT\_ID}\left(N-1,T_{Q^a},\Delta\mathrm{id}\right)$, $\mathbf{OPT\_ID}\left(N-2,T_{Q^a},\Delta\mathrm{id}\right)$,…, $\mathbf{OPT\_ID}\left(1,T_{Q^a},\Delta\mathrm{id}\right)$. To determine $\mathbf{OPT\_ID}\left(N,T_{Q^a},\Delta\mathrm{id}\right)$, not only is the solution of $\mathbf{OPT\_ID}\left(N-1,T_S,\Delta\mathrm{id}\right)$ required, but also $\mathbf{OPT\_ID}\left(N-1,T_{Q^a}-T_N,\Delta\mathrm{id}\right)$, the best solution for the first $N-1$ orders with the remaining trader sum $T_{Q^a}-T_N$, which constructs the constraint as $\left|\frac{\mathbf{OPT\_ID}\left(N-1,T_{Q^a}-T_N,\Delta\mathrm{id}\right)}{\left(T_{Q^a}-T_N\right)}\right| \leq \Delta\mathrm{id}$. The recursion can then be summarised in the following steps, if $Tran_N$ is not one of the transactions in the final subset S_N, then the transaction N can be ignored and $\mathbf{OPT\_ID}\left(N-1,T_{Q^a},\Delta\mathrm{id}\right)$ determined, however if $Tran_N$ is among the transactions, an optimal solution for the remaining transactions $1,\ldots,N-1$, must be sought, which is $\mathbf{OPT\_ID}\left(N-1,T_{Q^a}-T_N,\Delta\mathrm{id}\right)$. Using this set of sub-problems, the $\mathbf{OPT\_ID}\left(N,T_{Q^a},\Delta\mathrm{id}\right)$ can be expressed simply, in terms of values from the "smaller" problems, and so the recursion is summarised as two conditions:

1. If $Tran_i \notin S_N$, then $\mathrm{OPT\_ID}\left(N,T_{Q^a},\Delta\mathrm{id}\right) = \mathrm{OPT}\left(N-1,T_{Q^a},\Delta\mathrm{id}\right)$;
2. If $Tran_i \in S_N$, then $\mathrm{OPT\_ID}\left(N,T_{Q^a},\Delta\mathrm{id}\right) = T_N + \mathrm{OPT}\left(N-1,T_{Q^a}-T_N,\Delta\mathrm{id}\right)$.

Algorithm 1 is generated by a re-organisation of this recursive process, based on the two conditions above and can be applied by invoking $\mathrm{OPT\_ID}\left(N,T_{Q^a},\Delta\mathrm{id}\right)$ for N transactions and the capacity T_{Q^a}.

The Algorithm 1 gives, as has been shown, the S_N transactions from Q^a. The traders in S_N are potentially involved in collusive transactions. To further investigate, we need to identify the deviation degree of the transaction volume list discussed in step 3 in Section 3.3.1. If the volumes in the transactions in S_N are all similar, it indicates that the involved traders trade as passing a similar parcel along a certain loop within themselves. As discussed in Section 3.3.1, the identification is also tolerant of intended mixed transactions with different volumes. Therefore we use a straight forward approach for examining the transacted volume list. Once we obtain S_N, the transacted volume list can be represented as $\{V_1, V_2, \ldots V_N\}$. We first sort the list and ignore the top and bottom $\Delta\mathrm{v}_1$ extrema values respectively, where $\Delta\mathrm{v}_1$represent the percentage, i.e. 5%. We calculate the standard deviation V_{std} of the remained volumes and compare with the volume margin $\Delta\mathrm{v}_2$. If $V_{std} < \Delta\mathrm{v}_2$, we consider each transaction in S_N having similar volumes and thus the traders in $S_N$are involved in parcel-passing type collusive transactions. The identification approach is described in Algorithm 2.

Algorithm 1 Trader ID Matching Identification by recursion

1	ID_MATCH(Q^a)	// original transaction set Q^a;
2	$S_N = \emptyset$;	// solution subset, initialized to empty;
3	$T_{Q^a} = \sum_{i \in Q^a} \langle Tran_i \rangle_T$; N= number of transactions in Q^a	// T_{Q^a}: total trader number in Q^a;
4	OPT_ID$(N, T_{Q^a}, \Delta\text{id})$	// N decreases on each recursion step;
5	if $N < 1$ or $\left\lvert \frac{T_S}{T_{Q^a}} \right\rvert \leq \Delta\text{id}$	// if N reaches the last one or Δid condition is satisfied;
6	return;	
7	if $\left\lvert \frac{T_{Q^a} - T_N}{T_{Q^a}} \right\rvert \leq \Delta\text{id}$	// if condition is satisfied, then transactions in S_N is one solution;
8	output S_N;	
9	push $Tran_N$ into S_N	// assume $Tran_N \in S_N$;
10	OPT_ID$(N-1, T_{Q^a} - T_N, \Delta\text{id})$;	// recursively find solution by condition 2;
11	Discard $Tran_N$ from S_N	// assume $Tran_N \notin S_N$;
12	OPT_ID$(N-1, T_{Q^a}, \Delta\text{id})$;	// recursively find solution by condition 1;
13	end of OPT_ID	
14	return S_N;	

Algorithm 2 Collusive Transaction Identification

1	bool Collusive_Tran$(S_N, \Delta v_{1,2})$;	// original transaction set from Algorithm 1 S_N;
2	Sort(S_N);	// sort S_N in terms of the transaction volumes;
3	remove top and bottom Δv_1 transactions and obtain S'_N;	// remove the extrema;
4	Calculate the standard deviation of transacted volumes of S'_N: V_{std}	//
5	if $V_{std} < \Delta v_2$	// if standard deviation is within the margin,
6	return true;	// the S_N is potential collusive transactions
7	else	//
8	return false;	// if not, S_N is normal transactions
14	End	

4. Experimental evaluations

The evaluation of any detection model generally involves the use of genuine data from both genuine and abusive trading cases. However, because there is little tangible data arising from collusive trading, and in any case regulations forbid the disclosure of fraudulent market data, there is an extremely limited amount of genuine collusive trading data available, and certainly vastly less data than is available from records of routine trading. Thus, in order to evaluate the proposed detection model in a manner acceptable to the financial industry, in this study the characteristics and patterns found within collusive trading cycles are reproduced and introduced into original trading records, in order to generate a dataset that includes both normal and abusive trading cases (NANEX, Exploratory Trading in the eMini, 2013). An advantage of this is that such randomly synthesised manipulation cases can mimic any version of collusive trading; the collusive transactions can be generated at any time, with any volume sizes and margins. Furthermore, the use of synthetic financial data is already accepted in academia, for the evaluation of a proposed model, in the absence or dearth of real data (Palshikar & Apte, 2008) (Ho & Zhou, 2008) (Franke, Hoser, & Schröder, 2007).

For this study, the experimental evaluation occurs in two parts: part 1: experimental evaluation, using genuine trading datasets; part 2: experimental evaluation using genuine trading datasets with the addition of synthetically-generated wash trade scenarios, as per the analysis given in Section 2.1.

4.1 Experimental Setup

This evaluation uses genuine market data (in the form of transaction information) concerning four financial products, namely S&P 500 ETF (SPY), S&P 500 Index Future (Emini), 10-Year T-Note Futures (ZNM13), Eurodollar futures (GE13) from NYSE and CME group. Those products have been chosen because they were all reported to be targeted in manipulative collusive trading events in 2013. We obtain all collusive transaction data as well as the legitimate trading records in the same time period from our industry partners. The datasets cover full transaction records from May to June 2013 and comprise more than 1,000,000 transactions for each financial product. An excerpt of full information of the genuine S&P 500 ETF (SPY) data is shown in Table 4. The original dataset contains 14 items, which show a complete overview for every single transactions. Due to the customer confidentiality, the buyer and seller information (counterparty codes) are anonymized with distinguishable ID such as the client 1, 2, 5 and 7 in Table 4. The 'Instrument Code' and 'Instrument Description' items give the ID and one-sentence description of the traded security respectively. The 'Market ID' item shows which market the transaction was occurred in while the 'Order ID' item is the order sequential number in the trading system. 'Order Total Quantity' and 'Transaction Quantity Filled' represent the size of the first order and the final transaction. In the two transaction records in Table 4, the transacted sizes are equal to the first order sizes. That indicate the initially placed orders being completely filled. In some cases, the initially placed order (for example, a sell order with size 100) cannot be 100% filled (for example, 80 shares of the sell order can be filled by a later matched buy order). In that case, the 'Transaction Quantity Filled' will be smaller than the 'Order Total Quantity'. 'Limit Price' and 'Filled Price' items are the prices of initially placed orders and the transaction. 'Settlement Currency' item indicates the transaction currency. The 'Number of Fills' indicates the number of orders that were filled with the initially placed order. For example, if a sell order with size 1000 shares was initially placed in the market and two later orders with size 500 shares were then filled with the sell order, the 'Number of Fills' is equal to two. The last two items, 'Entered Date Time' and 'Last Execution Date Time' show the date and time of initially placed order and the transaction respectively. In our study, since the transaction size, prices and counterparties are considered crucial conditions for collusive clique recognition, the items, 'Counterparty Code', 'Transaction Quantity Filled', and 'Filled Price', are used in the experimental evaluations.

Table 4 An excerpt of full information of the S&P 500 ETF (SPY) data

Instrument Code	Instrument Description	Market Id	Counter-party1 (seller) Code	Counter-Party2 (Buyer) Code	Order ID	Order Total Quantity	Transaction Quantity Filled	Limit Price	Filled Price	Settlement Currency	Number of Fills	Entered Date Time	Last Execution Date time
SPY	SPDR S&P 500 ETF Trust	NASDAQ MAIN	Client 1	Client 5	00005497634 ORLO0	1270	1270	163.3	163.4	USD	1	20130503 09:34:37	20130503 09:35:56
SPY	SPDR S&P 500 ETF Trust	NASDAQ MAIN	Client 2	Client 7	00005497636 ORLO0	30000	30000	165	163	USD	1	20130503 09:35:01	20130503 09:36:12
……	……	……	……	……	……	……	……	……	……	……	……	……	……

Our proposed two detection approach: coarse and fine identification, are all applied to all of four datasets to identify firstly the trend of traders clusters and secondly any collusive transactions in a parcel passing format. Furthermore, synthesized collusive transactions as the examples in Table 1 are injected into three datasets for further evaluation.

4.2 Evaluation experiments using original datasets

Initially, the proposed approaches are evaluated on the four original datasets. Each of the dataset contains reported collusive transactions. On SPY, a big group of collusive traders were reported by NANEX while on other three datasets, several cases of parcel-passing collusive transactions were reported (NANEX, Chicago PMI, 2013). Thus, the evaluation is to reveal how applicable the proposed approaches is to real reported cases. For the experiment on SPY, we choose the parameter k=2 and 3 as the example in Figure 5, as long as one of the cluster identified reflects to the reported collusive traders, we consider the result is accurate. For the experiments on Emini, ZNM13, and GE13 datasets, we adopt straight forward measure true negative rate, TNR, and true positive rate, TPR, to evaluate the test results, where $TNR = \frac{TN}{TN+FP}$ and $TPR = \frac{TP}{FN+TP}$. The true positive, TP, is defined as legitimate cases detected as being legitimate and false negative, FN, is defined as normal cases detected as being collusive transactions. In those experiments, we choose different trader ID margin Δid from 0% to 20% with interval 2% and different standard deviation margin as 10% and 20% respectively.

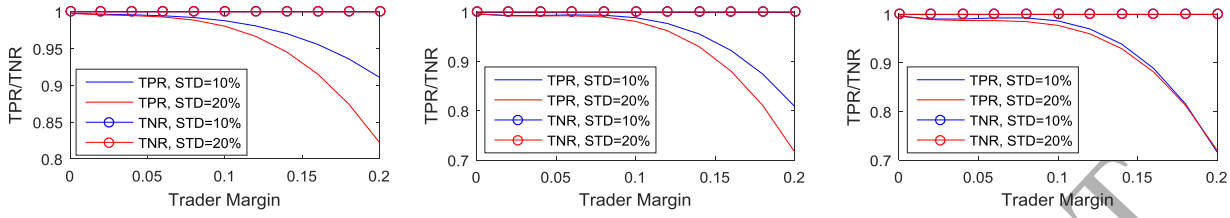
For the experiment on SPY dataset, a group of collusive traders are identified under both k=2 and 3. We find out that the group of traders identified by our approach contains 53 traders, more than the reported case, which only contain 40 trader accounts as the collusive parties. However, from the data we obtain, the 13 traders also trade

heavily among the collusive cluster. After further investigation and the consultation with our financial partner, we find out that the reported collusive clique was actually composed by 12 firms, which control a bunch of different trading accounts, among which, the reported 40 trader accounts are the most active ones but do not mean the collusive clique is limited to merely the 40 accounts. Those 13 traders we identified are actually accounts controlled by those 12 firms and also involved in the case. Therefore, we consider our approach effectively identify the collusive clique.

The experimental outcomes on Emini, ZNM13, and GE13 datasets are shown in Figure 6. The TPR and TNR measures under different trader ID and standard deviation margins are illustrated as curves in Figure 6. It is apparent that the TNR is always 100% under any configurations. It means that our approach effectively identify the collusive transactions in practice in all selected datasets and no manipulative cases have been identified as normal ones. The TPR, however, changes across different margins. It indicates that large trader ID margin brings more legitimate cases to be identified as collusive transactions. This result, in one hand, reflects our expectation that in a broad and flexible range, i.e. 20% trader ID and standard deviation margins, a group of traders who generally trade normally with a small part occasionally trading collusively in a parcel-passing style may be identified as collusive clique, on the other hand, still raises some suspicious cases. We have carefully verified some of the false negative outcomes, and consulted closely with our partners from the financial industry. As a result, it can be stated that the false negative cases detected are similar in many aspects to parcel-passing style activities, although they have not been picked up by regulators as such. Table 5 shows the transactions in two detected false negatives case. Case #1 shows trader Client1 selling 2239 shares to Client2 at a price of 1631.45. These were bought back on the next trading day with slightly lower price 1630.00. The transactions between Client 1 and Client 2 show almost identical transacted volumes, and a closed cycle of transaction directions, thus although unreported, this shows many of the signs characteristic of the parcel-passing action. The fact that the proposed approach has detected this is a demonstration of its effectiveness - however further verification of the nature of Case #1 requires regulatory input, and thus falls outside the scope of this paper. In Case #2, Client 1 sells 10239 shares to Client 5, the price being 1624.56. This occurs before market closing time, and Client 1 goes on to re-purchase all of the shares, at a slightly increased price 1630.60, when the market opens the next trading day. These transactions also fulfil the conditions of parcel-passing styled collusive transactions. Financial experts suggest that both Case #1 and 2 are the case of pre-arranged trading, whereby “a sell is coupled with a buy back at the same or pre-arranged price that limits the risks” (SEC, 2011). The key (indeed the only) difference between this and collusive transactions is that pre-arranged trading usually involves two parties and can occur over different days, whereas collusive transactions may involve multiple traders and is usually an intra-day occurrence. When the proposed detection method is used to identify collusive transactions specifically, it can be set to apply to intra-day transactions in order to avoid detecting cases of pre-arranged trading such as that in those cases. However, it is important to note that pre-arranged trading of this type is also illegal (SEC, 2011) and methods are required to identify it and eliminate it from capital markets.

Table 5 False Negative cases of Emini

case#	time	volume	price	seller	buyer
1	31/05/2013 15:30	2239	1631.45	Client1	Client2
	03/06/2013 09:12	2400	1630.00	Client2	Client1
2	31/05/2013 16:20	10239	1624.56	Client1	Client5
	03/06/2013 10:05	10239	1630.60	Client5	Client1



(a) (b) (c)

Figure 6 Test results on original dataset, (a) S&P 500 Index Future (Emini); (b) 10-Year T-Note Futures (ZNM13); (c) Eurodollar futures (GE13);

4.3 Experiments on datasets with synthesized wash trade data

The use of synthetic data in testing environments allows researchers to mimic any or all possible collusive transactions cases and to evaluate the robustness of proposed methods under any given collusive transactions scenario, such as random combinations of one or more traders in manipulative activities.

4.3.1 Generating wash trade cases

For the purposes of this evaluation, typical collusive transactions activities are reproduced and injected into each of the four datasets. These reproduced activities can be summarized as: a random number of traders trade along a random loop among themselves with transacted volumes similar but not identical (with random differences under certain ranges). To achieve a comprehensive assessment, the trader number in a clique is set in a range of 2 to 200. The loop sequence was randomly generated and the transacted volumes are fixed value plus random differences, which are set within 10% of the value. We keep the same configurations on the detection approaches: different trader ID margin Δid from 0% to 20% with interval 2% and different standard deviation margin as 10% and 20% respectively. The inherently random nature of this data allows the authors to thoroughly evaluate the proposed approach, using any possible collusive transactions scenario. The approaches are tested on four genuine financial datasets, each of which is injected with 1000 synthetic collusive transaction cases. Therefore in total 4000 completely random experiments are conducted, which constitutes a robust evaluation of the proposed model of detection.

The synthetic collusive transactions are combined with the real data of corresponding financial products, thus the test data comprises both legitimate and abusive patterns. The time intervals between pairs are random, as per the examples given in Table 1. Therefore the legitimate transactions might be mixed inside the synthetic transactions, which might occur in a genuine trading environment. Thus the experiment is conducted in circumstances that accurately reflect reality.

4.3.2 Experimental Results and Implications

Figure 7 shows TPR and TNR curves of the results of experimental values across stock datasets of SPY (Figure 7(a,b)), Emini (Figure 7(c,d)), ZNM13 (Figure 7(e,f)), and GE13 (Figure 7(g,h)) under different margin configurations.

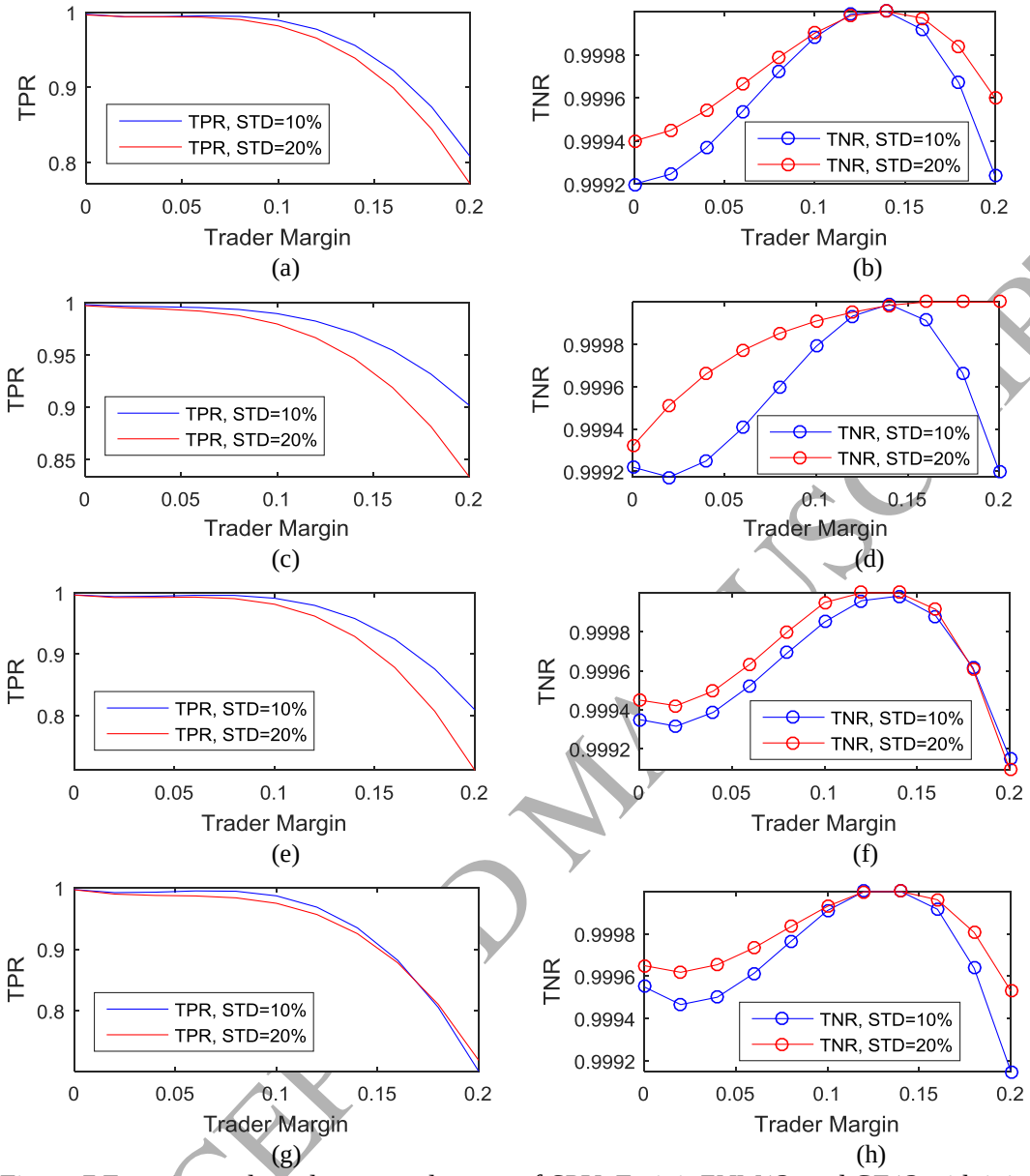


Figure 7 Experimental results across datasets of SPY, Emini, ZNM13, and GE13 with injected synthetic collusive transactions

From Figure 7(a,c,e,g), we can observe that TPR results are basically following the experiments on real dataset in Figure 6: the TPR value decreases in line with the increase of trader ID margin. However, the TNRs results show some variations. From Figure 7(b,d,f,h), we can observe that the TNR achieve the highest value100% when trader ID margin is around 12%-15%. Under other margins, the TNR rates decrease. This result follows our expectation that a greater number of potential collusions can be revealed under a relatively large margin value, which might be used to offsets "smart" trader's strategy. As the discussion in Section 3.3.1, "smart" manipulators do not follow a simple format of collusion. They trade collusively as well as normally. Identifying the collusive actions from the mixed trading records requires a flexible approach that can be tolerant to revised format of the manipulative activity.

A relatively large trader ID margin implies that the collusive clique with heavily trading among themselves can include a marginal percentage number of traders who might not trade with the clique as heavily as others but are unconsciously involved in some trading activities with the clique. This might be the case that the collusive orders among the clique are unintentionally picked up by other traders or the case that the traders inside the clique intentionally trade with others outside since the collusive trader can also trade normally to cover their collusive activities. However, a very large margin value, for example 20% in the example in Figure 7, achieved a relative lower TNR value. That indicates that including a large percentage number of traders who don't trade heavily when recognizing the collusive clique might heavily bring irrelevant traders so that immediately reducing the performance of recognising the transaction loop in Algorithm 1 and Algorithm 2. Consequently, either a small or a large trader ID margin does not bring the best recognition accuracy. Selecting the margin is a trade-off between compensating the 'smart-trader' tactic and accurately recognising the collusive trader loop. To achieve a better performance, selecting and tuning an appropriate value of trader ID margin based on historical data is crucial for each single security. In another word, the ability of the proposed method to configure the margin makes it particularly practical in a genuine trading environment, where each financial security requires a tailor-made configurations according to different market features. After tuning the margin based on historical data, as the experimental results, our approach can effectively and stably identify the collusive actions under the optimized configuration, i.e. 15% trader ID margin. Overall, the experimental results suggest that the proposed approach uncovers primary collusive transaction scenarios very effectively and consistently across all selected data.

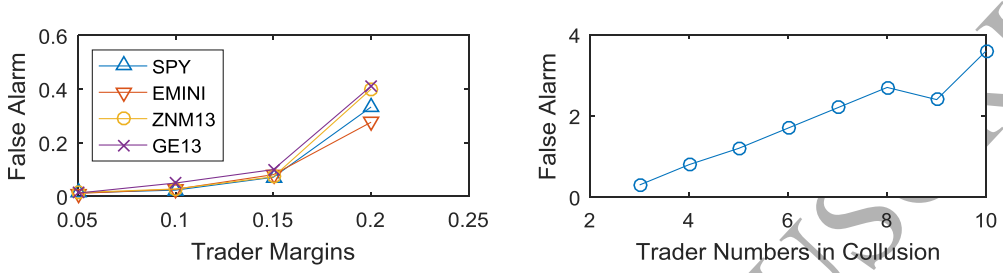
4.3.3 Theoretical Comparison

As discussed in Section 2.1, collusive party detection problem was also thoroughly studied in (Palshikar & Apte, 2008). In their study, each trader in a clique was assumed to trade with all other collusive traders and their crossed transactions were identified as a feature of collusive clique. This assumption is not true in most of cases in practice when the each trader only trade with parts of a big collusion. Our approach is not based on assumptions but only consider different margin values. In the experimental evaluation section of (Palshikar & Apte, 2008), only the detected collusive clique number (true negative) and the normal trading activities that were identified as collusive clique (false negative) were reported as the results. In addition, their experiments were all on synthetically generated data while our experiments were all on real market data. Although there are a number of differences between their study and ours, we still use their result as the bench mark for theoretical comparison as it was the only previous study on the identical problem.

4.3.3.1 Accuracy and false alarm

As reported in Table 5 in Section 6.3 in (Palshikar & Apte, 2008), collusion-clustering algorithm achieved 100% accuracy on detecting the collusion set under the best configuration. In our experiments, Figure 7 also reported 100% true negative rates on 15% margin across four different datasets, indicating that all collusive cliques were detected completely and stably in all experiments. In (Palshikar & Apte, 2008), the false alarm was used as the error measure of the normal trading activities identified as collusive clique. In Table 6 in (Palshikar

& Apte, 2008), collusion-clustering algorithm reported 0.3 false alarm as the best case for detecting a single case of 3-trader collusion. When the collusion contained ten traders, the false alarm increased to 3.6 for detecting a case of collusive clique. In Figure 8, we followed (Palshikar & Apte, 2008)'s measure, reported our false alarms, and compared with their results. Figure 8 clearly shows that the highest false alarm in our approach is around 0.4 at 20% trader margin and is slightly higher than the lowest false alarms in (Palshikar & Apte, 2008) algorithm, 0.3. In average, the false alarm of our approach is significantly lower than the algorithm in (Palshikar & Apte, 2008). Under the best configuration of 15% trader margin, our approach achieved the 100% accuracy at a low cost of 0.1 false alarm when detecting in real market data where the collusion included hundreds of traders. The algorithm in (Palshikar & Apte, 2008) can also achieve the 100% accuracy but with a much higher cost of 3.6 false alarm when detecting 10-trader collusion. Therefore, our approach significantly outperformed (Palshikar & Apte, 2008)'s algorithm on the false alarm.



(a) (b)

Figure 8 (a) false alarm under different trader margins for four datasets, SPY, EMINI, ZNM13, and GE13. (b) false alarm reported in Table 6 in **(Palshikar & Apte, 2008)**.

4.3.3.2 Complexity

In (Palshikar & Apte, 2008), the algorithm complexity was reported in two measures. In big O measure, the time complexity of the algorithm is $O(kn^3)$, where k is the number of nearest neighbours and n is the number of traders, while the space complexity is $O(n^3)$, and in physical time in real tests, the algorithm reported 1000 seconds for detection collusions in a dataset with 1000 traders. To compare with it, we reported our approach under the same measures. As the discussion in Algorithm 2, our approach is a recursive algorithm, which iterates each transaction for only once, therefore, the time complexity of our approach is $O(T)$, where the T is the total number of transactions. Since the recursive algorithm requires to temporarily save each iteration in the memory, the space complexity of our approach is $O(T!)$. The physical testing time across four different datasets are shown in Table 6. It is very clear that, although the two approaches are not comparable under big O measure, our approach is significantly faster than the algorithm in (Palshikar & Apte, 2008) even when applied on big datasets (more than a million transaction records across thousands of traders) and the performance is stable across four different datasets. However, according to the big O measure, our approach requires more memory in operation.

Table 6 Physical Test Time for four datasets

	SPY	EMINI	ZNM13	GE13
Test Time (seconds)	72.93	145.85	111.46	110.42
Number of Traders	1349	2396	3694	2986

Although our approach achieved a faster performance, we still believe that this approach is more appropriate for overnight screening in real financial environments. On one hand, the HFT traders can flood the trading system

by placing huge amounts of orders with later cancellation in daily trading. It's extremely challenging for any regulators to track the HFT in real-time due to their unbeatable speed achieved by super high performance computing facilities. On the other hand, regulatory inspection usually takes months for determining a manipulation case. The real-time monitoring would not shorten the process but brings huge cost on computing facilities. Consequently, compared with the algorithm in (Palshikar & Apte, 2008), our proposed approach can achieve high detection accuracy and low cost of false positive with faster detection time. But our approach requires more computer memory in operation due to high space complexity. When applied as an over-night screening tool, this drawback of our approach, we believe, can be compensated in real application due to the low memory costs.

5. Conclusion

This paper proposes a novel approach for the detection of collusive clique in financial markets. This method is suggested in light of careful and thorough study of various behaviours and scenarios in collusive clique. Analysis of the collusive activities undertaken is given in the form of directed graphs of traders, with transactions shown by directed connections between vertices. This illustrates the basic structure of collusion, which follows a closed cycle of transactions between given traders. Further work shows that in collusive clique, transactions are usually executed within a collusive traders with similar volumes. In light of this understanding, this paper propose two identification approaches for detecting collusive trends of the traders in a general format and the parcel-passing collusive transactions among the traders. To the best of our knowledge, the proposed method offers two key contributions to the expert system as well as the financial surveillance areas.

1. In the expert system area, well-known *k*-means clustering and unified dynamic programming algorithms are revised and effectively applied to collusive clique identification problem. It's the first time those two algorithms are thoroughly analysed and tailor-made to specifically apply in trading activity monitoring problem. The effectiveness of the application outperformed the previous studies in this area. Therefore this paper, on one hand, contributes to the expert system literatures on computational algorithm application in financial area, on the other hand, provides an inspiration that even a widely-used computational algorithm can be revised and applied to effectively solve a real and complex problem.
2. In the financial surveillance area, it is the first time the collusive clique detection problem is thoroughly analysed, extracted and formulated in a practical way and two approaches are inventively proposed to coarsely and finely solve this problem respectively. Through this, this paper provides effective tools for regulators according to their different detection requirements and also fills the gap in the literature by suggesting a novel approach to detecting a wide spectrum of collusive cliques.

Rather than restrict itself to detecting buy/sell orders of comparable prices, as in the "engine level" detection mechanism for CME, this proposed method allows users to identify collusive clique and their activities on a wider scale, taking into account any suspiciously transactions and collusive groups, and to assess these in light

of trading activities within a defined and sustained time period, in real time. Therefore, the authors believe that this approach is appropriate for overnight use in real financial environments and can be a powerful decision-support tool for the regulatory authorities.

However, our approach is still far from perfection. The frequency of trading is increasing rapidly, which poses a challenge to detection mechanisms in terms of efficiency. Real-time surveillance of the collusive clique is one of the future strand and might also be implemented in two ways: coarse and fine detection. For coarse detection, a sliding window mechanism can be applied with our proposed clustering approach to monitor the traders' trend in a pseudo real time. The fine detection can also be operated in every 2 to 3 hours in real trading environment. Based on our experimental results, screening the 3 hours' transaction data merely requires hundreds of seconds, which might not be difficult to be implemented in pseudo real time as the future work. In addition, as the transaction records and volumes in one single day increase rapidly, we might require a pre-screening method to remove the transaction records that are obviously not involved in collusive trading, for example transactions with unusually tiny or huge volumes. However, such pre-screening method requires more empirical studies of the financial market in the future. In the last, the collusive transaction format might be changing fast. To maintain the effectiveness of our approach, adaptive adjusting our approach according to the collusive format is also a crucial future strand.

Acknowledgement

This research is partially supported by the National Education Department of China, and the Fujian Province Nature and Science Foundation, P.R.China, project No.2015J01236.

Aitken, M., Harris, F. R., & Ji, S. (2009). Trade-based manipulation and market efficiency: a cross-market comparison. *22nd Australasian Finance and Banking Conference.*

Andonov, R., Poirriez, V., & Rajopadhye, S. (2000). Unbounded knapsack problem: Dynamic programming revisited. *European Journal of Operational Research, 123*(2), 394–407.

Bowen, C. (2013, July 9). *Market Regulation Advisory Notice.* Chicago Mercantile Exchange Group. Retrieved from http://www.cftc.gov/stellent/groups/public/@rulesandproducts/documents/ifdocs/rul070913cmecbotnymexcomandkc1.pdf

Cao, L., Ou, Y., & Yu, P. (2012). Coupled Behavior Analysis with Applications. *IEEE Transaction on Knowledge and Data Engeering, 24*(8), 1378-1392.

Cao, Li, Coleman, Belatreche, & McGinnity. (2015, Feb.). Adaptive Hidden Markov Model with Anomaly States for Price Manipulation Detection. *IEEE Transactions on Neural Networks and Learning Systems, 26*(2), 318 - 330.

Cao, Li, Coleman., Belatreche, & McGinnity. (2014). Detecting price manipulation in the financial market. *International Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, (pp. 77 - 84). London.

Cao, Y., Li, Y., Coleman, S., Belatreche, A., & McGinnity, M. (2013). A Hidden Markov Model with Abnormal States for Detecting Stock Price Manipulation. *2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC),*, (pp. 3014 - 3019). Manchester.

Cao, Y., Li, Y., Coleman, S., Belatreche, A., & McGinnity, M. (2015). Detecting Wash Trade in Financial Market Using Digraphs and Dynamic Programming. *IEEE Transactions on Neural Networks and Learning Systems*, available online.

CESR. (2011). *Market Abuse Directive.* Paris: The Committee of European Securities Regulators.

CME. (2014, August). Retrieved from U.S. Commodity Futures Trading Commision: http://www.cftc.gov/filings/orgrules/rule082814cmedcm001.pdf

Cumming, D. J., Zhan, F., & Aitken, M. J. (2012, September 12). *High Frequency Trading and End-of-Day Manipulation.* Social Science Research Network.

Franke, Hoser, & Schröder. (2008). On the Analysis of Irregular Stock Market Trading Behavior. In *Data Analysis, Machine Learning and Applications* (pp. 355-362). Springer-Verlag Berlin Heidelberg.

Franke, M., Hoser, B., & Schröder, J. (2007). On the Analysis of Irregular Stock Market Trading Behavior. *Data Analysis, Machine Learning and Applications.* Freiburg.

Franklin, & Douglas. (1992). Stock Price Manipulation. *The Review of Financial Studies, 5*(3), 503-529.

Franklin, Lubomir, & Mei. (2006). Large investors, price manipulation, and limits to arbitrage: An anatomy of market corners. *Review of Finance, 10*(4), 645-693.

FSA. (2006, March). *The Code of Market Conduct*. Retrieved from http://www.fsa.gov.uk/pubs/hb-releases/rel52/rel52mar.pdf

Hautsch, N., & Huang, R. (2011). *Limit Order Flow, Market Impact and Optimal Order Sizes: Evidence from NASDAQ TotalView-ITCH Data.* Social Science Research Network.

Hautsch, N., & Huang, R. (2012). The market impact of a limit order. *Journal of Economic Dynamics and Control, 36*(4), 501 - 522.

Ho, T. B., & Zhou, Z. H. (2008). Domain-Driven Local Exceptional Pattern Mining for Detecting Stock Price Manipulation. *PRICAI 2008: Trends in Artificial Intelligence.* Hanoi.

ITG. (2010). *Global Trading Cost Review.* Investment Technology Group.

Jamal, N. (2012, December). *LSE broker fined for 'wash trade'*. Retrieved from http://dawn.com/news/771335/lse-broker-fined-for-wash-trade

Jiang, Y., & Jiang, Z.-P. (2014). Robust Adaptive Dynamic Programming and Feedback Stabilization of Nonlinear Systems. *IEEE Transactions on Neural Networks and Learning Systems, 25*(5), 882 - 893.

Kleinberg, J., & Tardos, E. (2005). *Algorithm Design.* Addison-Wesley.

Lee, E. J., Eom, K. S., & Park, K. S. (2013). Microstructure-based manipulation: Strategic behavior and performance of spoofing traders. *Journal of Financial Markets, 16*(2), 227 - 252.

Loh, T., & Cumming, G. (2012). *Market Manipulation: Safe Harbour for wash trades and matched orders upheld.* Hong Kong: Hong Kong Securities and Futures Ordinance.

Menyah, K., & Paudyal, K. (2000). The components of bid-ask spreads on the London Stock Exchange. *Journal of Banking & Finance, 24*, 1767-1785.

NANEX. (2013, May 31). *Chicago PMI.* Retrieved from http://www.nanex.net/aqck2/4304.html

NANEX. (2013, July 10). *Exploratory Trading in the eMini*. (NANEX) Retrieved from http://www.nanex.net/aqck2/4136.html

Ni, Z., He, H., Wen, J., & Xu, X. (2013). Goal Representation Heuristic Dynamic Programming on Maze Navigation. *IEEE Transactions on Neural Networks and Learning Systems, 24*(12), 2038 - 2050.

Palshikar, & Bahulkar. (2000). Fuzzy temporal patterns for analysing stock market databases. *Proceedings of the international conference on advances in data management* (pp. 135–142). Pune, India: Tata-McGraw Hill.

Palshikar, G. K., & Apte, M. M. (2008). Collusion set detection using graph clustering. *Data Mining and Knowledge Discovery, 16*(2), 135-164.

Patterson, S., Strasburg, J., & Trindle, J. (2013, March). *'Wash Trades' Scrutinized.* (Wall Street Journal) Retrieved from http://online.wsj.com/article/SB10001424127887323639604578366491497070204.html

Poirriez, V., Yanev, N., & Andonov, R. (2009). A hybrid algorithm for the unbounded knapsack problem. *Discrete Optimization, 6*(1), 110–124.

SEC. (2011, 10). *US Securities & Exchange Commision*. Retrieved from http://www.sec.gov/answers/limit.htm

Tsang, E., Olsen, R., & Masry, S. (2013). A formalization of double auction market dynamics. *Quantitative Finance, 13*(7), 981-988.

Wang, J., Zhou, S., & Guan, J. (2012). Detecting potential collusive cliques in futures markets based on trading behaviors from real data. *Neurocomputing, 92*, 44 - 53.

Zhai, J., Cao, Y., Yao, Y., Ding, X., & Li, Y. (Sep 2016). Computational intelligent hybrid model for detecting disruptive trading activity. *Decision Support Systems*, 10.1016/j.dss.2016.09.003.

Zukerman, M., Jia, L., Neame, T., & Woeginger, G. J. (2001). A polynomially solvable special case of the unbounded knapsack problem. *Operations Research Letters, 29*(1), 13-16.